Egg Sales of a Sri Lankan Shop

Report for STATS 4A03 Final Project

Samantha Yang

April 2025

1 Introduction

The selected dataset details 30 years of egg sales for a small local shop in Sri Lanka [Karunarathna, 2023]. It was originally used for forecasting competition in 2023. The objective of this work was to forecast to predict future egg sales beyond the 30-year time series by using time series analysis methods. Such methods included selecting and fitting seasonal ARIMA models and performing model diagnostics. The data were recorded daily, from January 1st 1993, until December 31st, 2021. It contained two columns: one for the date, and another one for the raw number of sales stored as an integer.

In the context of the dataset, it is important for a business to anticipate their sales so that they can stock products adequately and be prepared to meet the needs of their clients, maximizing their profits. By modeling egg sales, we may connect this business dataset to real world events. For example, egg sales in this dataset were found to be seasonal, perhaps due to cultural reasons and holidays.

2 Modeling

To begin, the data quality was checked. No missing values existed. There were 18 days in March of 2020 were zero sales were recorded, presumably due to the COVID-19 pandemic. Zero values in the time series causes issues with carrying out a Box-Cox transformation due to taking logarithms of the variable, so using the function `na.approx` from the Zoo package, these zeroes were replaced by interpolation. Though the true number of sales during that time were likely to be zero, this step was taken just for modeling purposes and it would be unlikely to impact the model fit itself. Having these exceptionally low values in the data was an anomaly and should not be reflected in forecasting. For this work, only data starting in 2010 were used. Though we were provided 30 years of data, having a shorter time series is less difficult to interpret and may be faster to run code in R.

To verify the existence any periodic trends in the dataset, the spectral density of the data was estimated. The most prominent signal in the data occurred at approximately a one-year frequency. This means annual or 365-day cycles. This was done in Figure 1. The other notable peak in the spectral density plot occurred at 0.25 of a year, or quarterly. This indicates possible variation in egg sales within a single season or quarter, although the annual signal was far more prominent. We interpret this by concluding that a seasonal ARIMA is likely needed.

c

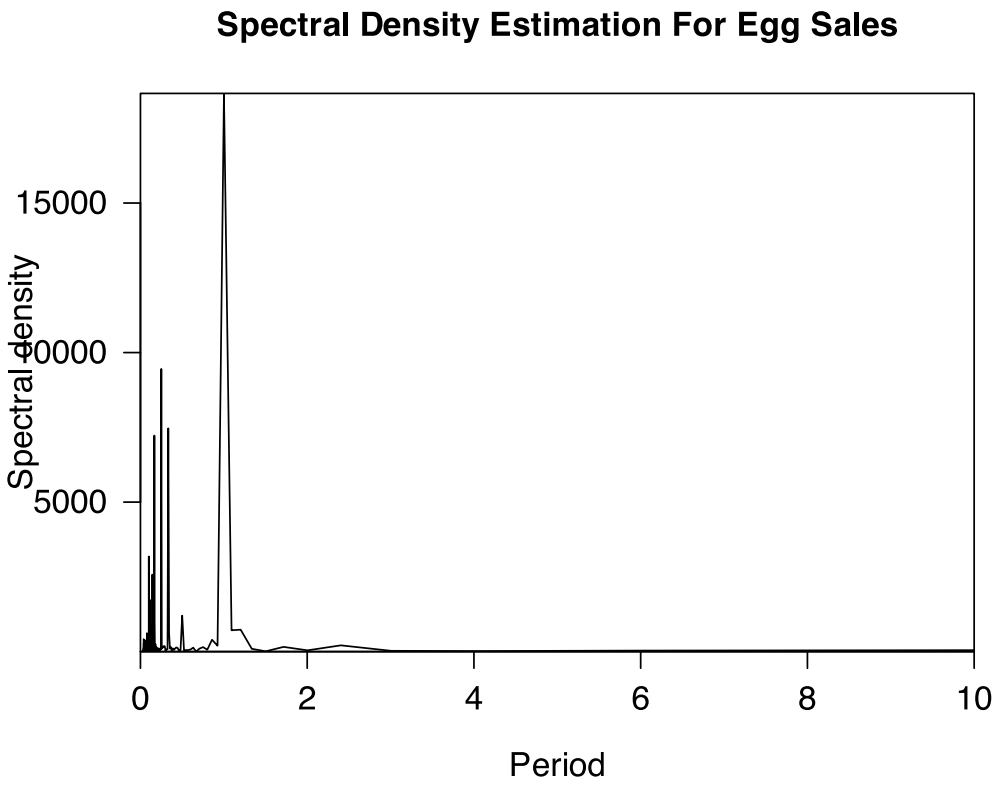

Figure 1: Spectral density plot for egg sales data

The `arima` function in R was unable to fit a seasonal model using a period of 365 days. This is a limitation imposed by compute power. Based on the spectral density, a seasonal model was very likely, so for the purpose of accurate seasonal model fitting, the daily data were transformed into weekly data by aggregating the sum of the sales over every week. The data were trimmed so that an exact number of weeks was included and that the first and last week consisted of data from full weeks. This way, the data are not misleading and no data points were too low. Thus, the transformed dataset began on January 4th, 1993 and ended on December 26th, 2021.

Then, the original time series plotted to get an idea of how it was distributed. There is an increasing linear trend with time, indicating that differencing was likely required. The data are also repeating on regular intervals, presumably annual. We also used a Dickey-Fuller test for stationarity. Its p-value was less than 0.01, so we reject the null hypothesis that the data are not stationary. The data are stationary. Next, using a Box-Cox transformation, we transformed the data as follows:

$$x = \frac{x^{\lambda} - 1}{\lambda} \tag{1}$$

The parameter λ was estimated using the function `BoxCox.ar` in R.

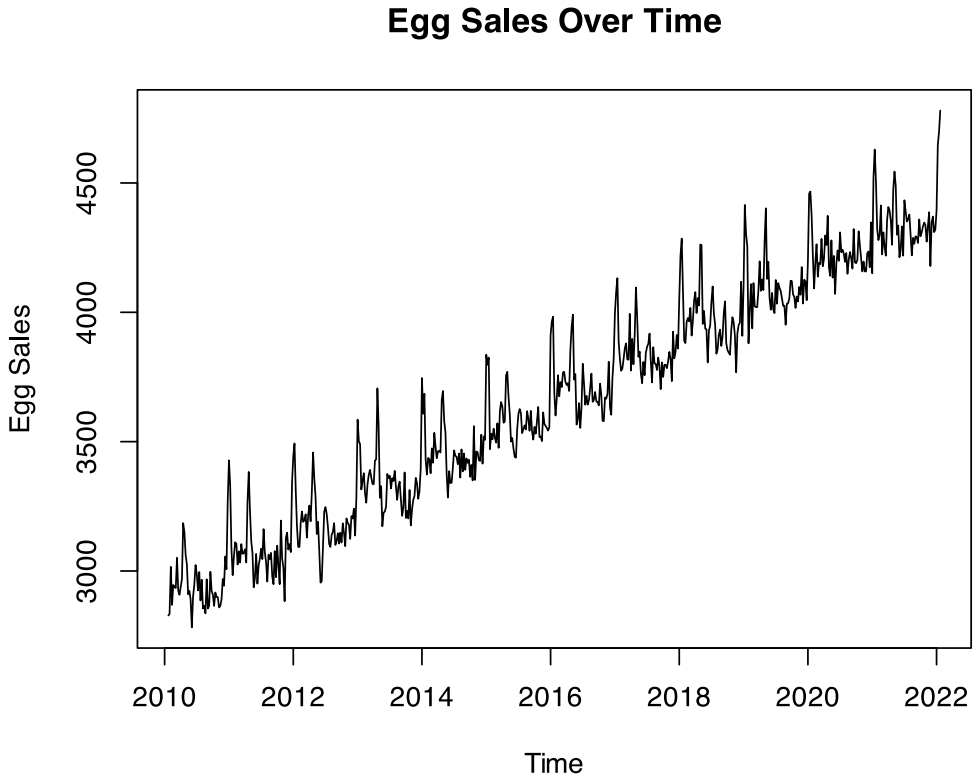

Figure 2: Time series for egg sales

The first step in selecting a model was plotting the autocorrelation function (ACF) of the transformed data. All autocorrelations were very high and outside the critical window, thus, the first differences were taken and the ACF and PACF of those differences plotted. This time, lag 1 had a significant autocorrelation and there were higher lags where autocorrelations fluctuated beyond the critical window, indicating that an MA(1) component would be included in the model, as well as a seasonal component. None of the partial autocorrelations of the first differences was signfificant; there is no AR component.

After taking the first seasonal difference in the model with a lag of 52 for the 52-week annual cycles, more information about the model was gained. These represent Figures 3 and 4. In this ACF, only lag 1 is significant; the seasonal component has an MA(1) component. The PACF was more complex, with significant autocorrelations at lag 1 and beyond.

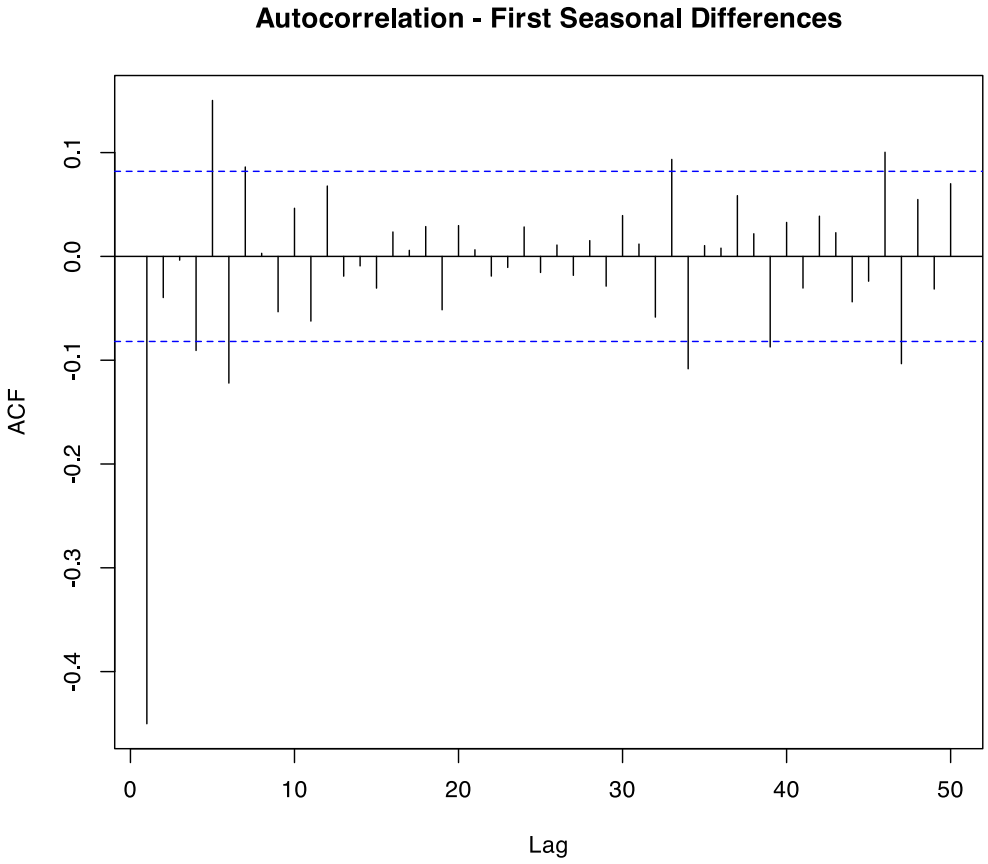

Figure 3: Autocorrelation for seasonal differences in egg sales data

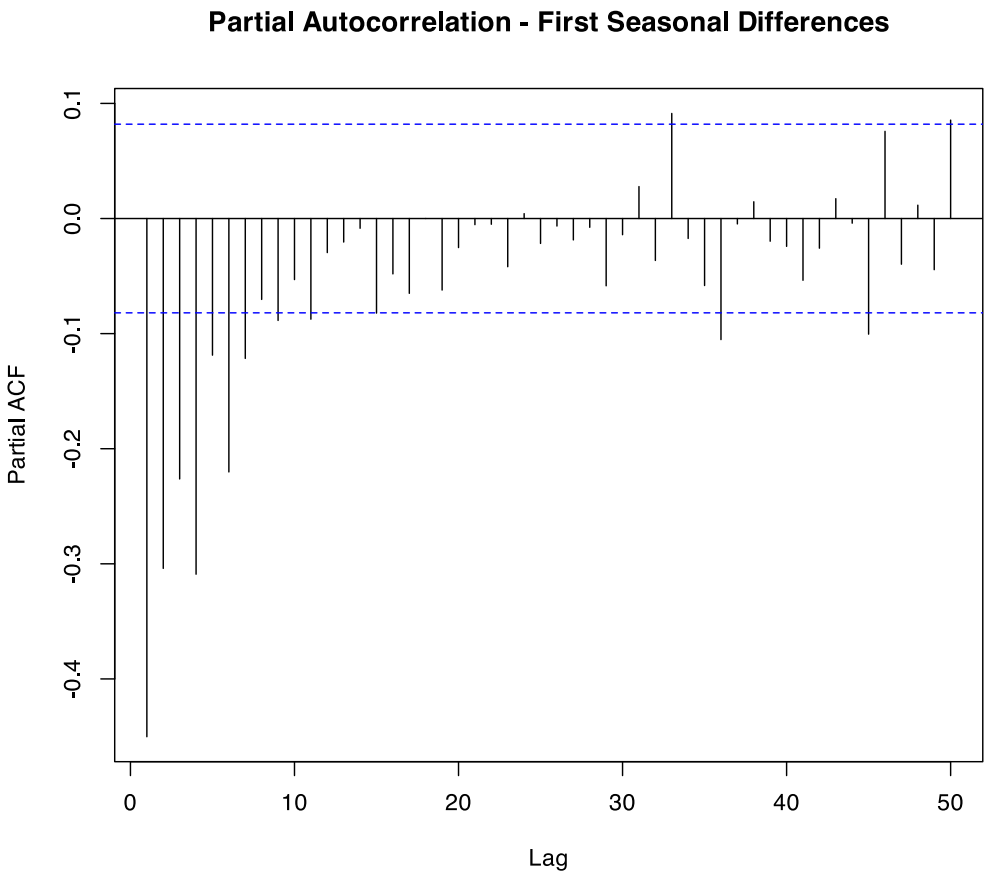

Figure 4: Partial autocorrelation for seasonal differences in egg sales data

Based on the ACF and PACF of the seasonal difference time series, a model with a specification of $\texttt{ARIMA}(0,1,1) \times (1,1,1)_{52}$ should be specified.

Due to computational complexity, models with higher AR components in the seasonal part were not fitted, for example, $\texttt{ARIMA}(0,1,1) \times (4,1,1)_{52}$. Because the time series was relatively long and it was computationally expensive to fit these models, it was not possible to run the line of R code in a reasonable time frame. The first partial autocorrelation was the most significant in the seasonal differences, so this model should be adequate. Using the function `arima` in R, we fit the model to get the following result:

```
   TSA::arima(x = egg_trans, order = c(0, 1, 1), seasonal = list(order = c(1, 1, 1), period
= 52))

Coefficients:
ma1  sar1  sma1
-0.9796  -0.1442  -0.4156
s.e.  0.0085  0.0804  0.0769
```

```
sigma^2 estimated as 0.04975:  log likelihood = 36.26, aic = -66.52
```

The residuals were analyzed using a few different methods. The residual QQ plot for normality was checked in Figure 5. The points fall mostly linearly except for some slight curvature. Typically, it is concerning if there is curvature in the middle of the points in the plot, not just at the ends. Whether or not this was concerning was verified using a Shapiro-Wilk test. This test has a null hypothesis that the residuals are not normally distributed, and an alternate hypothesis that they are normal. This statistical test returned a p-value of 0.4176. Since the p-value was greater than 0.05, we reject the null hypothesis. The residuals are in fact normal and so, there is no concern about the model in this respect.

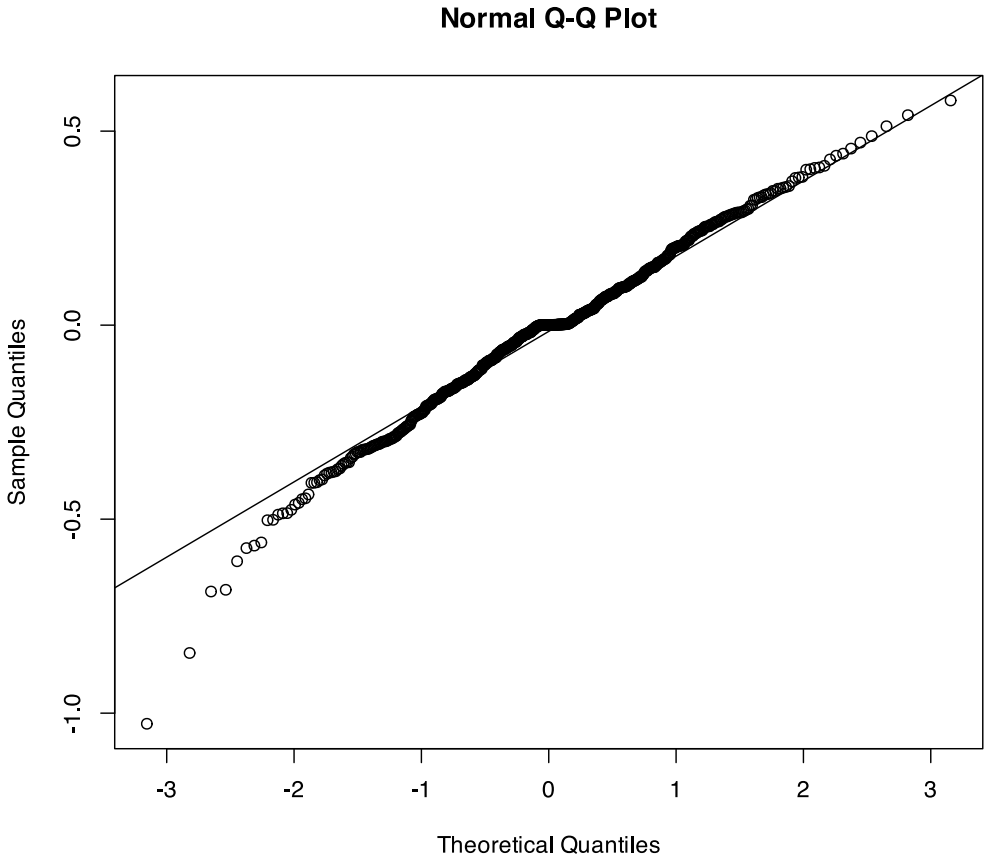

Figure 5: QQ plot for normality

The other diagnostics were also checked: the standard residuals' distribution, ACF of the residuals, and the Ljung-Box test. This was included in Figure 6. Looking at the standardized residuals, they appear to resemble white noise. The ACF of the residuals and the p-values for the Ljung-Box statistic show that there is no correlation between residuals, which is a good indicator that the model supports the data well.

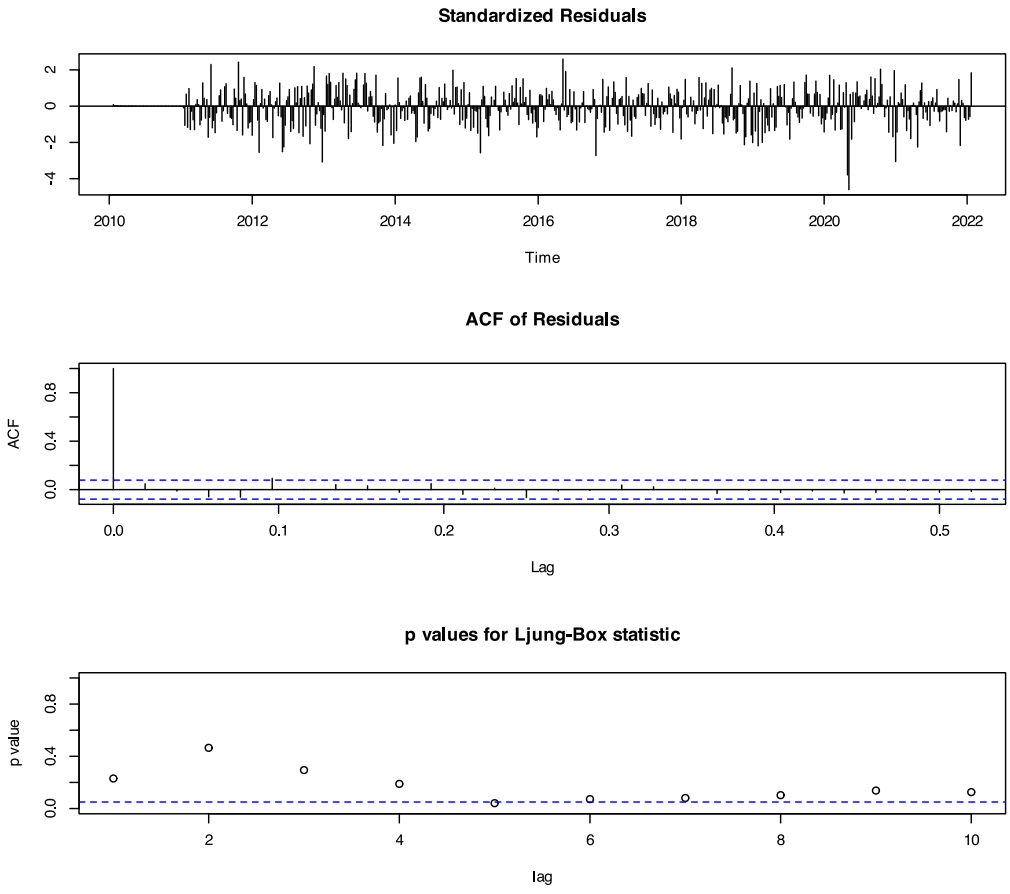

Figure 6: Diagnostics for SARIMA model

Thus, none of the residual diagnostics look concerning and we may conclude that the data do indeed support the SARIMA model with specification $(0, 1, 1) \times (1, 1, 1)_{52}$. Finally, we used the SARIMA model to forecast 52 time points into the future, equivalent to one year. This leads us into the Results section next.

3 Results

The prediction of the model was plotted in Figure 7. There are no data for egg sales published beyond the end of 2021 and egg sales are subject to fluctuations due to business constraints and social factors, as well as world events. However, due to the effectiveness of the model in capturing the trends of these data over approximately 11 years, we may conclude that these results are reliable. Looking at the figure with the predictions, the forecasts into future weeks resembles how the data look in 2021, so the model was a good fit.

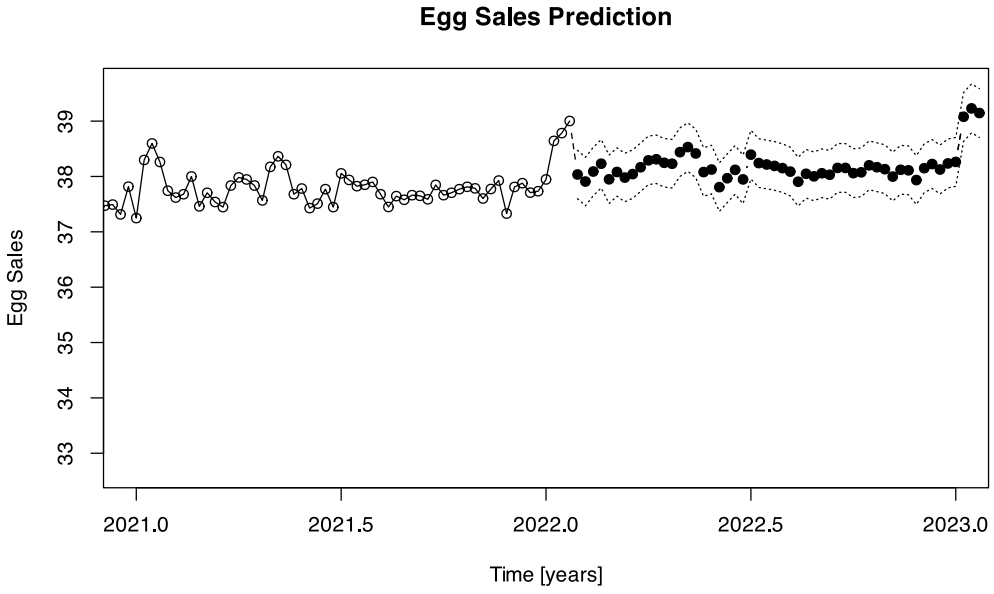

Figure 7: SARIMA model and predictions: the points within the original time series are outlined with no fill while the points of the prediction are coloured black with their intervals included.

4 Conclusion

For future work, the shorter periods with strong signals in the periodogram may be considered. It is possible that egg sales may be periodic by quarter or even by month, though the strongest repeating cycle in the dataset was annual. Our model accounts only for the annual cycles but it may be interesting to account for these other patterns in the model. Then, perhaps more precise forecasts for shorter time scales into the future may be possible.

Furthermore, this study was limited by computational power. It would have been useful to take higher order AR components in the seasonal part of the model for overdifferencing to much higher orders, perhaps up to AR(4) for the seasonal component. With a more advanced computer, it would have been good to experiment further with different model specifications and to carry out the diagnostics on those models to see if there would be any improvement. However, the current model was proven to be effective for this dataset and was able to get reasonable looking predictions. In the real world, these models are instrumental for successfully running a business.

References

[Karunarathna, 2023] Karunarathna, K. (2023). Egg sales of a local shop for 30 years.